Technically Optimistic

An Emerson Collective Podcast

In the final episode of our limited series on AI, we look at the big issues of accountability and responsibility. How should we allocate the responsibilities for managing this technology? Who will decide when AIs are doing more harm than good? Will we be looking to private companies or depending on public servants? And what will be left for individual citizens to decide?

***

RAFFI VO:
I'm Raffi Krikorian, and this is Technically Optimistic. It's episode six… and it's actually the final episode in our miniseries all about artificial intelligence.

We've talked about the history of AI, the risks, and the drive for regulation. We've talked about using AI in the classroom, and the importance of teaching people about AI. And we've talked about some potential impacts that AI will have on art, culture, and the economy.

But… there have also been a few themes that have sort of… kept popping up. There's the issue of whether AI will be used to augment humans, or to automate and replace humans. Then there's the daunting and complicated question of whether we need to revise our understanding… of what it means to be human.

But there is another big theme that has come up again and again in my conversations. And that is about responsibility… and accountability.

We're already living in the age of AI. And I propose that two things are true: First, the future is going to feature more and more AI in our lives.

And second, we do NOT have it all figured out yet! And I don't just mean technically… I mean from a legal, social, political, and most of all… ethical perspective. We have far more questions than answers on all those fronts.

How should we allocate the responsibilities for managing this technology? Who will decide when AIs are doing more harm than good? And what institutions will have the responsibility of jumping in to correct things? Private companies? Or to public servants? That's a pressing issue here in the US… where it seems like our government and big tech companies constantly point fingers at each other for who's responsible.

And for those of us who didn't even ASK for artificial intelligence in our lives… What responsibilities fall to each of us?... And wait– before you say "NONE AT ALL!"... keep in mind that we are not just individuals. We are community members. … Citizens. … And we gotta work together…. for the public good.

So, how should we sort all this out? And how will we know if we get it right? We're gonna give it our best shot… because we… are Technically Optimistic.

MUSIC…

RAFFI VO:
It can be hard to get our bearings around questions of responsibility and AI…. because of how fast everything's developing.

MARIA RESSA:
So in the old days it was, it was actually at a pace of human change. You could manage it still,

RAFFI VO:
Maria Ressa has a long view of the interplay between technology and society. With a particular focus on the truth.

RESSA:
And then now as you're talking about it, We're shifting all the time, right?

RAFFI VO:
Ressa is a Filipina journalist, who reported stories on terrorism and extremism in Southeast Asia for almost thirty years. In 2012, she co-founded Rappler, a news organization designed to produce social change. It covered the authoritarian tactics of Rodrigo Duterte, who became President of the Philippines in 2016.

CLIP:
VOICE: Nicknamed "the punisher," Rodrigo Duterte is a Harley-riding, former city mayor who speaks his mind.
DUTERTE: I am a president of a sovereign state
VOICE: Duterte came to power on an anti-establishment ticket, tapping in to strong disillusionment with those in power.

RAFFI VO:
Ressa exposed corruption, violence, and propaganda, shining a light on how Duterte's regime used social media — Facebook in particular — to try and manipulate the Filipino people.

CLIP:
RESSA: The Philippines is Facebook country: 97% of Filipinos on the Internet are on Facebook and when we came out with the propaganda series in early October of 2016 I was pummeled into silence, although I still talk… (fades out)

RAFFI VO:

The government… retaliated.

CLIP:
VOICE: A court in the Philippines has convicted a prominent journalist of the crime of cyber libel.
VOICE: It's being seen as a blow to media freedom in her country. Ressa's new cite Rappler is known for its tough scrutiny of President Rodrigo Duterte.
VOICE: Ressa now faces a prison term of up to six years.

RAFFI VO:
Facing prosecution, criminal charges and jail time… Ressa didn't back down. She spoke out… on the importance of free speech and the press, and about the dangers of new media technology.

CLIP:
RESSA: Uh, I've been shot at and threatened, but never this kind of death by a thousand cuts.

RAFFI VO:
And then, in 2021, she won the Nobel Peace Prize.

CLIP:
VOICE: I congratulate Maria Ressa and Dmitry Muratov on being awarded the 2021 Nobel peace prize. Throughout the world, the free press is essential for peace, justice, sustainable development and human rights.

RAFFI VO:
And though Duterte left office in 2022, Maria continues to use her platform to raise awareness on the dangers of technology,,, calling for accountability.

CLIP:
RESSA: *Our greatest need today is to transform that hate and violence the toxic sludge that's coursing through our information ecosystem prioritized by American internet companies that make more money by spreading that hate and triggering the worst in us.*

RAFFI VO:
Maria Ressa is a personal hero of mine. And when it comes to thinking through tech and responsibility… and how to prioritize our values in a free and democratic society… her perspective is invaluable.

RESSA:
Frankly in the old world I lived in, where there is corporate responsibility, until you know what it's going to do or how it is going to impact people who use it, you don't roll it out. OpenAI, why

would they roll out ChatGPT, they rolled it out in November. Why would like a grownup company like Microsoft do that? But now they're doing it right? But I think part of the reason is because they need people to test it at a scale that you could not do in a controlled environment. So that's a risk.

RAFFI VO:
One of the things I've noticed is that you use very specific language to talk about these things. In *How To Stand Up To A Dictator*, for those who don't know, that's your book from 2022, you're not just talking about autocrats… you call the big tech companies "dictators." And when it comes to social media algorithms, you have specifically said: don't use the term "models," call them "clones." You have to tell me more about that strategy. What are you trying to do by using and emphasizing these specific words?

RESSA:
Safiya Noble actually wrote about the flattening of meaning. Corporations used to do this in the past, to um, to, it's like a corporate communications thing. But the problem is when tech does it, and they are not transparent.They're not transparent about the dangers to us. So for example: model. Building a model sounds…

RAFFI:
Innocuous…

RESSA:
Innocuous. That's the right word. I do communications. I've spent my entire career. Learning how to tell stories from remote parts of the world.I spent my entire career for almost 20 years learning to take what is happening here in front of me in different languages and different cultures, catching your attention over there, which is a different language and a different culture, and making it fit, because I live in my countries. I'm held accountable by the people in the countries I live in. So my stories, because it's seen by the world – and this is what tech doesn't do – because we all see the same thing, I am held accountable for here in Manila and I'm held accountable wherever else you read it and see it, but I have to put it into language you understand. I very specifically use words translated into words that the people I am communicating with understand, right? Because in the end, why do I say clone, that they clone us? Because they literally do. A clone is a different version of you, that is exactly what they really have done. But they won't say it that way because then they'd have to ask us for permission. And then they would have to pay us. I guess part of, part of what I'm trying to do is restore meaning to the flattening of meaning,

RAFFI VO:
The idea of meaning… is gonna be an important thing to keep in mind, as we head into the future. That's what Maria is talking about, holding tech companies to meaningful descriptions of what they're up to. There's also the idea of meaning, as opposed to

nonsense… and making sure we can tell the two apart, especially as more and more AI-generated material makes its way into our information ecosystem.

Then there's the meaning that can only come from encountering another human being… not just from the words they say, but from the presence of inner, emotional life.

ROSALIND PICARD:
Affective computing was originally defined as computing that relates to arises from or deliberately influences emotion.

RAFFI VO:
Rosalind Picard is a scientist, inventor, and researcher at MIT's Media Lab, and is a pioneer in developing the relationship between computers and human minds. Her work in affective computing has led to, among other things, some of the first biometric wearable devices ever adopted.

According to Roz, if we want computers to interact naturally with us, we must give computers the ability to recognize and understand human emotions. Or even… the ability to express emotions themselves.

So, what does all this have to do with AI?

PICARD:
There are a lot of ways machine learning ties in, and some of them are very subtle. For example, most machine learning gets in, has a human in the loop, giving them some reward or punishment, some correction, some error signal. We say that what it did was right or wrong. The human says that. The machine has no morals. And so the human giving it that feedback, in a sense, is giving it a positive indicator or a negative indicator. A human getting that feedback would see that as an affective signal. You're telling me, I did well or not. I tend to feel good when I get the positive signal. I tend to feel a little bad when I get the negative signal. So we associate feelings with them as people. The machine, of course, has no feelings. It just gets these positive or negative signals. So at the very core of machine learning, there is this notion of getting positive and negative feedback.

So initially, The focus of Affective Computing was to show more honor and respect for human affect, to say, "hey, human beings, we're not just doing math and language and other intelligence tasks. We are fully embodied people who have feelings." And if you're gonna show respect for a human, you need to look at the whole package. So the idea was anytime there's machine learning or AI optimization in that interaction, we should be showing respect for the feeling part of that by recognizing it and by treating it the way that we would want it to be treated if there were a human in the loop interacting with that. Now that's very different from trying to build an AI psychiatrist. The growth in mental health problems, the inability to serve the need out there has increased the demand for technology that could augment what people are doing there. Today, the AI really cannot replace an expert human therapist. Psychiatrists, psychologists are needed now more than ever, human ones. That said, there is a lot more that AI can do today than it could do even last year. For example, the chatbots are trained to take a lot of conversational elements that a human expert would say if you said a certain thing to it, like “I'm really depressed, what should I do to help myself?” It is now much better trained to say something that a human would say in that situation, because it's been shown tons of examples of what are good things to say in that case. Does it know what it's saying? No. Does it know that it's a good thing? It only knows it got a positive reward mechanism, right? That that's considered an accurate response, but it doesn't know your feelings, it doesn't care about you, it doesn't know what matters to you. It is really a smart parrot that has been taught a collection of things that would be okay it's been told to say in response to what you say.

There is a lot of richness that humans pick up on when we assess the context of a situation before delivering care and support. And interestingly, that's one of the key things AI is clueless about. It's trained on text, for example, but it's not trained on context. It is context-free usually, and when it's given, all of these different... text. There are stylistic context things that it's getting better at. But it has to be told explicitly.

RAFFI VO:
So one of the dangers we've been talking to people about is that humans already have this incredible ability to anthropomorphize things. And now, these things are gonna have the power of language, which is going to make it even easier for us to think of them as human. So, do companies have a responsibility to think about how the AIs they’re creating will affect humans on an emotional level? Or from a mental health perspective?

PICARD:
I was just talking to some EU regulators about this. We were talking about what if companies were actually responsible for the mental health of their users, or at least for not harming the mental health of their users, right? Taking ultimate responsibility for it sounds like too high of a calling. But if you're actually harming people, that... sounds like something we should be preventing. And how would that be shown? And what would be an intervention for that? So this do no harm that is already the ethic in medicine… Now that we can more easily measure mental health and changes in it, perhaps we should be monitoring the impact on communities of the technology as well. So this is something that is actively being looked at right now, and it's likely I'd say to start first in the EU because they're much more aggressive in this space. But I would hope in the US we would try to own caring for our users, not just in the... data security sense, which is very important, and we need to make improvements there, but in the honoring and respecting their health sense as well.

MUSIC…

RAFFI VO:
It would be a good thing, if, somewhere in the process of AI development, someone took some responsibility for the health and wellbeing of users. But as you heard Roz say… it seems like

too high a goal for tech companies. And, sadly, that makes sense. With incentive structures aimed toward revenue and growth, tech companies have other things they're worried about.

JAMES MANYIKA:
AI is such an exciting... technology and it's so foundational to everything, quite frankly, in the sense that I can't imagine any... activity by individuals, by organizations, by companies, and even governments where we won't be harnessing the capabilities and the potential of this technology.

RAFFI VO:
James Manyika is the Senior Vice President of Research, Technology, and Society at Google.

MANYIKA:
And I think, you know, with that comes two things as with any powerful transformative technology, all the amazing possibilities. but also some risks and challenges that we're going to need to navigate. So we're going to have to somehow hold these two things together and quite frankly work our way through both of those things. On the one hand, how do we make sure that we harness all the potential and also in particular harness it for everybody as a public good?

RAFFI VO:
Manyika.. recognizes the limitations in our current incentive structure…

MANYIKA:
That won't happen by itself. Commercial interests alone won't get us there.

RAFFI VO:
But he also wants us to be clear-eyed and rigorous when discussing AI risks. And that's important… for questions of responsibility.

MANYIKA:
On the other hand, we have to think through the complexities and risks. And I think one of the things I think is important on the complexities and risks is to be thoughtful first of all about the different categories of those risks and challenges.

They're not all the same. One is the set of risks that I think of as the systems aren't doing what we'd like them to do. They're either giving us biased outputs, toxic outputs, inaccurate outputs, No one intends that, but they're not delivering and performing. with the outcomes we would like. So there's those kinds of risks. And I would actually put things like bias in that, and things like factuality…

RAFFI
Okay.

MANYIKA:

…in that, and toxicity. I don't think anybody designs these to be toxic. Then I think you've got a second category of risks, which is different, which is what I think of as misapplication and misuse. So even if the systems work the way you want, you could imagine entities, whether they're individuals or organizations or companies, or even governments, misapplying or misusing these systems to do other things that none of us want. So that's where you might have somebody deliberately using it to create disinformation, for example, somebody deliberately using generative AI to make up falsehoods. So in other words people misapplying or misusing in ways that cause harms to people in society and I think it's worth even their making decisions between misapplication and misuse. Misapplication may be somebody just using this in a way that didn't realize it was going to do that. They apply, it works here, so they also apply it over here. It wasn't actually designed over here.

RAFFI:
Oh yeah.

MANYIKA:
I think of that as misapplication. It can be benign in the sense that no one, unintended. Then there's also actual misuse. Somebody's actually deliberately misusing these systems, either for commercial, criminal, political, whatever reason, they're misusing the system. So you've got that category. Then I think there's a third category, Raffi, where the risks are, think of these as second-order effects. Right, so we use these to get innovation and productivity. Oh, by the way, we're starting to impact jobs as a consequence of it, or we're starting to impact something in the labor market, or we use these and people start to start to harm mental health or something. People are overusing these. Think of those kind of as unintended consequences, if you like…

RAFFI:
Mm-hmm.

MANYIKA:
…or second order effects or derivative effects. Then finally, of course, you've got the risks, the fact that the systems themselves start to become very powerful and it's what people think of as kind of the safety control problem in the sense of we now have created systems that are more capable than we are or more intelligent than we are. Can we be sure that they're aligned with our values or you know can we control them etc etc. So I think it's worth thinking through each of these different categories of challenges and how we navigate our way through all of that.

RAFFI:
Can I ask how you feel responsibility lies on those risks? Is the responsibility with those people who are creating them because they didn't foresee the risks on like the, both the first order and those second order risks. Like who is responsible for them?

MANYIKA:
Well, I think everybody, let me explain that. Everybody in the following sense. I think it's worth thinking through the chain here, right? There's a bit of a chain that's emerging. You've got at one

end the developers of the technology itself, the underlying technology with its large language models or any of these systems. I think... they have responsibility to make sure the systems are performing as intended. SThen you've got people who are doing, if you like, who are further refining them. Then you've got people who are deploying the systems. They may not be the same as ones who've built them or refined them. Then also you've got users, right? Because users could take, it's kind of like the classic issue with, if I make a knife. It depends, well, are you going to use it to eat food or to kill somebody? So I think we have to think about the whole chain. Take what happens in the labor market. I can design a technology that complements what humans do. You on the other hand may choose to apply it to substitute for what humans do. It's the same technology. I may design a technology that's very useful for writing an essay or trying to understand a subject, but if you try to use it to get medical advice, that's a misapplication, right? So

RAFFI:
Yeah.

MANYIKA:
That's why even in a world right now where we're dealing with these large language models, there's certainly certain kinds of uses I wouldn't put one of these systems to. I don't think I'd get legal advice from a large language model. But would I use it to write an essay? Sure. I think when it comes to responsibility, we all... have to think about this collectively I think.

RAFFI:
I'm hesitant to push back too hard, but…

MANYIKA:
Oh, please do, please do.

RAFFI:
But one of your phrases reminds me of “guns don't kill people, people kill people.” Is that the analogy that you're trying to draw?

MANYIKA:
No, what I'm trying to describe is, you know, if I'm developing a powerful technology, this is why I said the developers have the responsibility. I need to make sure I'm making it as safe as possible.

RAFFI:
Mm-hmm.

MANYIKA:
Just like if I become a manufacturer, I need to test it. Does it meet society's expectations about safety? Do the brakes work? Will they stop the car when I need it to stop? Right. So I have to, as a developer of the technology, they bear some responsibility. Now, if I choose to take that car

and drive it into a crowded marketplace, how do I think about that? That's why I think. these questions about uses and use cases and users. We may even say, well, if the user, is the user qualified to use it? I don't know if that's a useful thing,

RAFFI:
Oh, that's interesting.

MANYIKA:
Right? It may be that what you might give a child. to use, versus what you might give a scientist to use, might be very different. If a kid comes into a laboratory with powerful chemicals, I think which cabinets I lock up are very different than the cabinets I would lock up if a scientist walked into that same lab with powerful chemicals.

RAFFI:
Yeah.

MANYIKA:
I think developers of technology who are computer scientists need to be aware of these ethical considerations. And now that's not to say all of a sudden have become expert ethicists.
But I think the people, the collection of people working on these systems should include these incredibly multidisciplinary backgrounds. And we try to do this at Google, right? Where we have computer scientists, but we have ethicists, we have philosophers, we have social scientists. I think the collective enterprise of developing these systems, all of these disciplinary expertise.
But at the same time Raffi I think as much as we spend a lot of time thinking about the risks and challenges. To me, the idea of getting AI right involves holding two things at the same time that fully, maximally benefit society and address the risks and challenges.

RAFFI:
Mm-hmm.

MANYIKA:
So we have to... have both of these conversations, not just the one side of them. And by the way, I think there's a fruitful tension between these two. One of the things that we're trying to do at Google is this idea of being bold and responsible. And I realize there's a tension there, but boy, it's a wonderfully productive tension. And I think it should guide what we do. And I often think that even when we're talking to policymakers and even regulators, this is a powerful enough technology, we should regulate it. We should do that with an eye towards both enabling fully and maximally the beneficial things to benefit everybody and also managing the risk. It's got to be both.

MUSIC…

RAFFI VO:

Google is uniquely positioned to live in that tension that James called out… between AI's risks and its benefits. Being both bold and responsible.

But who is primarily responsible for doing the research that will move AI forward? For most of its history, AI research has been conducted primarily at universities, as is the case for most fields where new knowledge is generated. But in recent years, many people have noticed a shift.. in where these studies are being conducted… Away from academia, and toward industry.

KYUNGHYUN CHO:
In fact, this change is nothing too new. That tide has been always moving toward or pointing toward industry.

RAFFI VO:
Kyunghyun Cho is a professor of computer science and data science at New York University. But also…

CHO:
I'm spending 50% of my time at the moment at Genentech,

RAFFI VO:
… which is a biotech company doing cutting edge work… using machine learning to design new antibodies.

CHO:
The reason being that a lot of let's say things that we do in AI or more broadly computer science are very, very directly relevant to the products. Now some technologies tends to be closer than the others to the actual product. So for instance, networking protocols or the networking technologies. They're extremely close to the real life products. We're using our laptops, we're using our mobile phones. All these things, if you just trace back 20, 30, 40 years ago, were only studied in academia. We were doing research, not building products. I think AI is going through something very similar. In fact, a lot of the algorithms that we are using in order to build this kind of large-scale language models or the vision systems, as well as self-driving cars and whatnot, many of those algorithms are not the algorithms that we designed or that come up with last year or two…

RAFFI:
Mm-hmm.

CHO:
…years ago or three years ago. Most recent algorithms that I can think of as the building block that goes into all these technology or the products is from let's say 10 years ago. And then if you think about the algorithm that we use for optimization that was in fact first described in 1950s in an academic paper. So it actually took a long time but what is now is that because everything has gotten so much closer to the actual products, the companies are just increasingly putting

more investment in this particular direction. That just gives us an illusion that there may not have been anything before. And suddenly, there's something, all these new things that are coming out of these companies, although it's mostly about how to productize them, and then kind of put it into the actual products that are feasible to the regular people, non-scientists, let's put it like that.

RAFFI:
But the incentive mechanisms of companies and academia are very different from each other. I mean, again, you wear both hats.

CHO:
Yes.

RAFFI:
I guess what I'm trying to struggle with is, it's hard to say good and bad, but are those incentives well-purposed? Is there a correct balance between the academic incentives and the commercial incentives?

CHO:
People like me and the people like my colleagues at academia, and in particular the ones who have been working on natural language processing for the past, let's say, 10 years or so, I think we've been somewhat too lazy. And then we were just enjoying the fact that we were in a field that was very clearly going through the huge transition. And then that was getting huge amount of traction. And that was actually getting closer and closer to the real products to the point that the industry was pouring money and resources into this field. And then what that means is that we really didn't have to think too far into the future, nor we have to innovate dramatically ourselves in order to stay afloat in this field that is natural language processing and more broadly artificial intelligence. So what that means is that our incentive in academia should always have been the curiosity, and then thinking about the future. So what the future is going to be like and what kind of science we have to do now in order to prepare ourselves for the 10 years, 20 years, if not 100 years in the future. But then unfortunately we're lazy because there was a lot of interest from the whole society, a lot of funding, and a lot of students wanting to join our lab and so on. Indeed the incentive structures do differ between academia and industry. But in this particular case during that time, this past decade, somehow this line between the incentive structure in industry and incentive structure in academia was so blurred, we started to get confused ourselves.

RAFFI:
You know, one of the superpowers of academia is that like, you don't have to be financially driven. So it was the laziness there that like, all of a sudden, like they've started looking at the same problems.

CHO:
Yeah, so it's not only finance, but more like the, you know, one of the things that we academics

all kind of crave for is the attention and…

RAFFI:
Hmm (laughs)...

CHO:
… popularity, reputation. And then it was so much easier for us to build up this kind of reputation or increase our visibility by working on this particular thing that was just expanding exponentially You know, it was a, I mean, I was lazy myself. So yeah, I'm kind of doing self-confession here. And that kind of, I think blinded a lot of, let's say us to the point that we couldn't really see that the, well immediately at this point any company would come out from the shade and they say that well here's the actual product that was built out of all those technologies that have been built during the past, let's say, decade or so. But as soon as that happens and then if the product is successful, that set of technologies become something that needs to be pursued in industry, not anymore in academia. And then we never actually kind of prepared ourselves for it.

RAFFI:
So then what do you need to do about it?

CHO:
We need to really think about science. So there is a really nice position paper that was posted on Archive just recently from a bunch of people from DeepMind, then Decoheer and whatnot. And one of the authors is my good friend Phil Blunson. It's about the scientific debt in natural language processing. And then what they point out is that a lot of papers that have been published over the past few years are so focused on simply showing that whatever the system they have built works ever so slightly better than the existing systems on a predefined set of the benchmarks, to the point that no one really knows what was the actual scientific contribution that was made from each of those papers.

RAFFI:
Okay.

CHO:
And then in fact, when you go deep into it and then run some kind of, let's say, extensive evaluation of the existing work or studies, then what you notice is that a lot of things that were indeed impactful, were not actually discussed as the actual impact that they were making in the original papers.

RAFFI VO:
The paper that Kyunghyun is talking about introduces the notion of “scientific debt.” Basically, the authors claim, a great deal of research on training large language models fails to do certain things that scientific research is supposed to do… that they’ve stopped explaining exactly how… their results were achieved.

CHO:
we've been chasing all this kind of fame and reputation and interest and the visibility to the point that the, our papers were really just a white papers of the products or the rather sloppily created products that we made out of the academia rather than providing the actual scientific insights as well as the scientific knowledge to the society. So what I think is that we just need tostop pretend to build a product that…

RAFFI:
Mm-hmm.

CHO:
We don't know actually how to build and then start doing science which is the actual product that we were supposed to and we were trained to produce.

RAFFI:
This, I mean, this sounds like a big deal. This sounds

CHO:
Oh this is a big deal, yes.

RAFFI:
Yeah. I mean, like you're basically telling your entire field of just like, we've been playing the wrong game.

CHO:
I think it's actually the whole society. So the incentive structures... that apply to, let's say, engineers who work at companies to build products, or the business people who are thinking about how to build products and sell the products, and then the incentive structure for the university to train students in a way that they're going to be very useful to the society and also are going to thrive after they graduate, and then incentive structure for researchers to think about science and whatnot. Now, for instance, let's think about the funding structure: Everything is about how good proposals are written by these professors, and then how viable their research looks like. But..

RAFFI:
Mm-hmm.

CHO:
…then the thing is, how can you actually decide on the future of what kind of things these people are going to do, and how much future impact this research will have? It's really difficult. Then you have to, what do, let's say, people do, is that they are going to punt that decision or the assessment to the others. But if you're working at the very frontier of a particular field, the group of the people who are working on that is very small. If everyone is actually in conflict with each other, then you start asking questions. Who are actually reviewing these proposals? And

then at the end of the day, how if you can... If you have to impress the external panel with your potential scientific plan to get the funding, the sure way is to use whatever is being touted as popular, reputable, or interesting at that point. That is why everyone is writing about the exactly same thing at the moment, language models, how to use them. I'm pretty sure there are hundreds of the proposals that are being written as we speak about how to use language model for X, Y, and Z.

RAFFI:
Sure.

CHO:
and then I tell those people and then tell myself as well because I'm actually one of those people is that they, well. If that is what we are going to do, it looks like these are exactly the products that we're building. Why are we actually doing it in university? We gotta just go out there, make a company, make actual products, right?

RAFFI:
This is exactly the conversation I wanted to have, just sort of understanding like how these incentives are playing off each other. Because like a lot of the people we've been talking to will say things like, well, companies are incentivized to move incredibly quickly. Whereas academics are supposed to be curiosity driven. But like what I hear you saying is like, they're actually on a collision course with each other. We've somehow like mixed it all together accidentally.

CHO:
Yeah, it was a happy accident, 10 years ago or 15 years ago, when we started to work on this kind of machine learning and then graduating from our grad schools and whatnot. Because back then, companies started to realize that, well, there is potentially a really big thing coming that may come out of this technology called deep learning. Even 10 years ago no one was teaching artificial neural networks. to their undergrads. No university was doing so.

RAFFI:
Hmm.

CHO:
What that means is that the companies realized that, well, this may be a big thing. We need to hire people, we need to invest in it, but there are only PhDs, very small number of PhDs. So people like me and my colleagues and my friends, we benefited a lot. And then that was the beginning of this kind of collision. And then this actually collision happened eventually. We didn't know that it was going to happen this fast, but it actually eventually happened. And people in industry? They are doing well and then you are actually in the right place because they have all those skill sets, they have the expertise and experience. Now they can really build all these products and then they are the ones who actually created or made this collision possible. Now people like me in academia, we should have done better and then thought about, okay, if there

is this inevitable collision coming in our way, what should be the next things that we need to prepare ourselves for? Because all these collisions of course happened during the past 10 years, but based on the technologies and the science that has been that had been done over the past half a century. You know, in academia we should have been essentially in the mindset of the people in the 1950s…

RAFFI:
Hmm.
CHO:
But we were in the mindset of the 2010s unfortunately. That's my bad.

RAFFI VO:
In the face of the incredible power and influence of the tech companies… who's gonna somehow have both the credibility.. and the means… to push back, if needed?

RAFFI VO:
Here's Maria Ressa again.

RESSA:
in 2016 when we rolled out the weaponization of the internet, a three-part series, One of the things we, we got clobbered, that's when I received 90, about an average of 90 hate messages per hour.

But they also, the pro Duterte, um, accounts, the, the kind of info ops they were working on began a campaign against us, which was enabled by Facebook, by social media. Right. It was called hashtag unfollow Rappler. R

RAFFI VO:
Mmm-hmm.

RESSA:
In about a month's time, they not only attacked us, attacked me, attacked at a very visceral level. They were tearing down credibility. This was all done. That month-long campaign not only served to try to tear down our credibility, try to tear down my credibility, begin to dehumanize me. What it did is; They were able to get 50,000 accounts to unfollow Rappler. Which at that point was about 1% of the number of followers we had. Right? I could quantify that in money terms…

RAFFI VO:
Sure.

RESSA:

So what does that mean? I'm saying that these things are harmful and for the tech companies to allow it. To allow information warfare, insidious manipulation, using our clones that they say they own. I think this is unconscionable. I wouldn't have used that word in 2016, by the way. I would, I didn't know.

RAFFI:

So one of my hypotheses, and I think I think you'll agree with this, but please tell me if you don't, is that tech is like another way to wield power. Like power is moved from like politicians to money to tech in some way. How, how should regular people then position themselves?

Like, should we actually democratize it? So should more people get access? Do we need more technologists? Do we need to expand tech education? How do we both push back but also maybe take back some of that power to, to wield it ourselves?

RESSA:
I don't think tech is just power, right? E.O. Wilson said this tech is God. It is godlike.

RAFFI VO:
We'll be right back… after a quick break.

MUSIC…

MIDROLL

MUSIC…

RAFFI VO:
Welcome back to Technically Optimistic, I'm Raffi Krikorian, and this episode we are talking all about AI and responsibility.

Maria Ressa's reporting on social media shaped her view of the internet. For her, it's a place desperately in need of accountability… for things like misinformation, abuse, and the subversion of democracy.

So, how did this crisis of responsibility come about? Who exactly… dropped the ball?

RESSA:
Actually I say two groups abdicated responsibility for protecting the public, for protecting us, you know, and I would say first it's tech. In the past I've compared tech to big tobacco. Because how much money is enough?

RAFFI:

Yeah.

RESSA:
So the first is tech abdicated responsibility. Who's the other group that abdicated responsibility? Governments. And that’s part of the reason I couldn't be a journalist in the old way. So I started focusing on how do we solve this problem? In the end, I keep saying that the public debate, especially in the United States for many, many years, centered on content moderation, which is not the problem. Facebook deflected to content moderation, you know, the oversight board, which is going to be like a Supreme Court for content. And I was like, I'm just a a, a good person because I didn't lash out at that point to say it isn't the content that's a problem, it's the distribution that's the problem. Right?

RAFFI:
Yeah.

I mean the beginning of the cascading failure, which is, you know, if you think about the information ecosystem as a river, content moderation is like taking a glass of water from the river, cleaning up the glass of water and then dumping it back into the river, right? So you have to go further upstream: what is polluting the river. And that is the factory of lies. And it's not just doing that like it is used by power players, domestic and international politicians or, or power structures to modify our behavior. It has become a behavior modification system. Like in many ways what the tech companies succeeded in doing is saying that the virtual world is a different world, but it really isn't because there is only one person that lives in both worlds. This is not about content, really what we need to do is to pull up to a, a bird's eye view of data flows. 'cause this is about data.

MUSIC…

RAFFI: VO:
How do you get this “bird’s eye view” that Maria is talking about here?

The Stanford Internet Observatory… or SIO… is an interdisciplinary, collaborative research network studying abuse of the internet.. in real time. SIO researchers analyze their observations and make policy recommendations to governments, and social media platforms.

DIRESTA:
SIO’s work on generative AI writ large, meaning just the capacity for machine-generated content. This is actually fairly new. You used to have to be a little bit more sophisticated…

RAFFI: VO:
That’s Renee DiResta, the research manager of the SIO.

DIRESTA:
We've seen text be democratized over perhaps the last year, but at SIO we had access to open

AIs, GPT-3 workspace for academics. So we had some access to this back in 2020, I think was when we started working on text generation.

RAFFI:
How does SIO technically detect what's going on? Like how do you actually decide this is a piece of content that we need to like thread through the system? Should I envision this as a big siren goes off?

DIRESTA:
A lot of the concern about these things being used in disinformation campaigns really requires vast distribution. But if you want to do something that's going to gain significant pickup or capture significant attention, this is something that you would have to do if you were trying to have an economic impact on a major company or really create a massive scandal around a particular, you know, ahead of an election or something like that. You do need accounts to distribute the stuff. And that’s where you do still see social media companies having some power here, right? Because one thing that their integrity teams are good at is identifying networks of fake accounts or accounts that all of a sudden are coordinating in a particular way to push out the same piece of content perhaps or to all simultaneously boost the account that has just put out that content. So you can use this technology to try to engineer a sensational moment, create panic by putting something out on a platform like Twitter and you see blue check accounts, this happened actually I think maybe a month ago, blue check at the Pentagon with some very badly generated, AI-generated imagery, and there's a bit of a panic as people try to figure out what is happening there.

> **CLIP:**
> *VOICE: Not long after the markets opened this morning, a picture -- a disturbing picture – and claims suddenly appeared on twitter. It showed what looked like an explosion stating it occurred near the Pentagon. In just moments the image started to panic, ricocheting across twitter and social media. But the thing is, it's all fake. It was a complete hoax, and once authorities began making clear the explosion was not real, the markets bounced back. The most disturbing maybe part about all of this: it wasn't sophisticated. Simple AI generated a fake image that sparked about a $100 billion stock market move. And this is nothing compared to the sophisticated AI manipulation and deep fake technology that is already on the market.*

DIRESTA:
When I saw that image go up that purported Pentagon attack, I think I saw it about 20 minutes after it dropped. The pillars were wrong, right? The physics of the building were wrong, the fence was wrong, that is not how fences stand, you know. If you look at it, if you just glance, you might be potentially taken in, but if you tell people, like, these models get the following things wrong a lot of the time, you can at least give them a little bit of a way to be informed customers, recognizing that three to six months out, those tells are no longer going to work. And so it is kind of, I think, a constant way to both say, here's the specifics, but also more importantly, here's the healthy skepticism that you should have when a truly sensational image lands in front of you. In

this moment in time, we can create these things very, very easily and we can make them very, very realistic looking and highly plausible and you just need to understand that is happening.

RAFFI VO:
So, what can we do to equip everyday people with the tools to not get duped by this kind of thing? Is it just some form of AI or media literacy? Like, how do you teach your son what he should be looking for when he goes online?

DIRESTA:
… So in a way, I have been explaining to him that the internet is fake. The internet is fake and propaganda is everywhere. That's what you hear about when you live in my house.

RAFFI:
Hehehehe

DIRESTA:
You know, and so here is how this tool is profoundly powerful. Here are all the very, very cool things we can do with it. And then we do talk about things like the ethics, right? Of when a machine is trained to replicate an artist's style, what does that mean for the artist right? So I just kind of walked them through like this is what is possible in the world and people try to trick you sometimes.

RAFFI:
But like the Internet isn't all fake. I mean, like I understand having a healthy level of skepticism, but like.. I mean, you're one of the experts on disinformation. How do we train our kids to understand how to flirt with that line of like, it is useful in the following ways, but it is fake in the following ways?

DIRESTA:
I think a lot of it is what is the intent of the speaker? What is the intent of the poster here? Why have they communicated this to you? Why did they make this picture? Why did they write this article? I actually pulled out the old archives for the Institute for Propaganda Analysis. It was in the 1930s, it was kind of the interwar period, and there was a bunch of academics who decided that they needed to deal with a particular influencer by the name of Father Coughlin, who was a preacher and had very anti-Semitic and kind of fascist rhetoric that he used. and he had about 30 million listeners on the radio at a time when there were 120 million people in the US, right? So just massive, massive reach.

> **CLIP**
> *FATHER COUGHLIN: There is no need of communizing all the factories and the fields and the forests and the mines under a new kind of God made of Flesh and Blood and clay and hatred… (cheering)*

DIRESTA:

Institute for Propaganda Analysis was an effort to try to explain to the public what the rhetorical techniques were. Why does this rhetoric work on you? Why does this message appeal to you? Why is this resonant? And that was related very specifically to a type of media literacy that was encouraging skepticism of everything, which just creates a sense of helplessness, right? A sense of Oh My God, the world is full of bullshit and how can I possibly filter through it? And that in some ways is one of the strategies that we see authoritarian governments using, right? Nothing is true and everything is possible. So how do you help people understand the certain types of techniques and why certain content facilitates those techniques and why they work on you without telling them, like, you know, question everything as if, you know, there's conspiracies around every corner?

MUSIC…

RAFFI VO:
Maria Ressa knows a thing or two about this particular strategy of authoritarian governments.

RESSA:
Where is the critical time? Between now and the end of 2024. At the beginning of the year, I started looking at how many elections are there between now and 2024. There are 90 key elections. And some of them have already happened, but critical because 60% of the world is under authoritarian rule today. This tech, the pipelines that connect us, because it makes facts debatable, because it spreads lies six times faster because it triggers fear, anger, and hate. Then we are electing illiberal leaders democratically all around the world. So last year, 60% of the world was under authoritarian rule. This year, V-Dem in January said it's 72%.

RAFFI:
Yeeks.

RESSA:
Where is the tipping point? Where is the critical time? Between now and the end of 2024. So what does that mean? You know, you asked earlier, is it gonna get worse? Of course. Uh, because, because profit is winning. The private tech company incentives determine behavior right now.

RAFFI:
Yeah I agree.

RESSA:
And the tech guys haven't behaved in a responsible manner, so, okay. Then they'll say, yes, but our primary responsibility, shareholder value. Okay. Well like, enlightened self-interest. Do you really want to, to kill democracy? They really should answer that question. A toaster has more safety measures and safety regulations to pass, uh, before it can get into your home than this thing that you carry with you everywhere you go that monitors your heartbeat. The every step you take that when you point in actually creates a clone of you, right? Like this is extremely

powerful. And I, for a very long time with Rappler, I saw the power for good. But I think as the tech companies in the United States began to realize what they had, I think they behaved like teenagers.

RAFFI:
Hmmm.
RESSA:
They were gleeful about it. I mean, you know, in the book I talk about how many engineers in Facebook had access to our data, which engineers were using it for their dating lives? How is that possible in a grownup corporation, right? Like this is standards and ethics is what we would call it in news, but corporate ethics, corporate responsibility, right? Like, You have to grow up and, and, and join the real world and be responsible for your actions. And all of this, the beginning of the cascading failures for an extinction event is information.You know, it's the arrogance of tech. I always felt, look at the beginning when tech. I mean, Mark, I, I had a conversation with him where he said, you know, well, you know, tech can do better. Tech can do better than politicians 'cause politicians and governance is messy. Democracy is messy, but really, tech can really do better? Well, that hasn't been our experience right now.

RAFFI VO:
As we've talked about before, there's an important tension between technology… and government. Tech wants to move fast… but government moves slow. Government, at least in theory, should protect all interests, of both private citizens and private corporations.

Tech companies… have no such mandate. Tech CEOs love to talk about improving people's lives… and of course many people working in tech are earnestly trying to do… just that.

But many of us have come to think of tech… and not government… as the force that will protect the public good. And… whether that's right or wrong… who's responsible for that story?

MEREDITH BROUSSARD:
There's this kind of fantasy around algorithms. There's this fantasy around using computers where people imagine that there's some kind of bright technological future around the corner, that someday we're going to make enough technology that it's going to deliver us from the essential problems of being human. And I definitely used to think that.

RAFFI VO:
That's NYU professor and data journalist Meredith Broussard. We heard from her a bit in episode 2, where she talked about how biases can be embedded in AI training data. It's… more than a glitch, she says… which is also the title of her newest book.

BROUSSARD:
I mean, that's how I was taught. I was taught to believe in this sleek, algorithmically me diated future. And it just kept being kind of crappy. Like, every future that got invented, it was like, oh, this is not necessarily better than what we had before. And again, you can only have that

happen so many times before you get disillusioned and you realize, like, wait, there must be some other kind of problem that is happening. And so the way that I characterize it is it's a kind of bias that I call techno-chauvinism. The idea that computational solutions are superior to others. And what I would argue instead is that we should use the right tool for the task. And sometimes the right tool for the task is a computer, absolutely, 100% yes. And then sometimes the right tool for the task is something simple like a book in the hands of a child sitting on a parent's lap. You know, one is not inherently better than the other. It's about, again, the right tool for the task.

RAFFI VO:
Is this technochauvinist viewpoint related at all to… the demographics of who's working in the tech industry?

BROUSSARD:
So when I was an undergraduate studying computer science at Harvard, I was one of only six women, right? And I only knew a couple of them, like I couldn't find the other ones. The diversity problem in Silicon Valley, it's still there.I have been... seeing these problems with tech for my entire professional life. And I've also been hearing the same promises about a bright technological future for decades now. And you can only hear the same promises so many times before you stop believing them. One of the things that has been really fascinating to me is discovering the extent of unconscious bias. and realizing that we all have unconscious bias. We're all working every day to become better people, but we're not there yet. We're not perfect. None of us is perfect. And so what happens is we all embed our unconscious biases in the technology that we create. And so when you have small and homogeneous groups of people creating technology, then the technology gets the collective unconscious bias of its creators.

RAFFI:
So the AR optimists seem to paint this picture of inevitability. Like, this is going to happen, so therefore we need to react against it. And I'm just curious if... if you too are hearing that, that inevitability argument going on and what can we do to actually then try to take back control of like: this is our society, not your society?

BROUSSARD:
You have just put your finger on the on the beating heart of my work. Right. I think that it starts with pushing back against techno chauvinism. I think it starts with acknowledging that there are many, many possible futures that we do not have to be beholden to the single technologically mediated future that has been dreamt up by a small group of Silicon Valley millionaires. And there is room in the world for so many different futures, so I want to empower people to understand technology so that they can push back against technological decisions that are unjust or unfair. Because right now people feel really, like there is the sense of inevitability of like when an algorithm makes a decision, it's like, oh yeah, you have to sit back and take it. Well, that's absolutely not true.

MUSIC…

RAFFI:
Maria, I know that this is obviously a huge question but… who can we trust to help us out of this mess? If the tech companies and the governments have both abdicated their responsibilities… where do we turn?

RESSA:
In the short term, it is just us, right? So what did we do in the Philippines? We did a whole of society approach, a four layer pyramid to protect the facts. We called it hashtag Facts First PH. And even though we have another President Marcos Jr. today, we literally were able to take over the center of the Facebook information ecosystem with facts. It became an influencer marketing campaign for facts. The Philippines is actually in a relatively better place because we are recovering rule of law. So in the short term I think the critical part is for every person on any of these things, To stop being a consumer, to stop being a user and to become a citizen, right? You need to hold these accountable and that is our task. And, and frankly, Americans have more power than most of the rest of the world because ~~it is,~~ these are American tech companies. You have systems that have been degraded, institutions that have been degraded, but you still can. I'm actually shocked that the US government allowed large language models out. You know, you have a strike going on right now about creation, 'cause the reality is that this is just computational processing of words, of language, looking for patterns in a very fast way. And I think in tech we're at the beginning stages of it, but if we don't survive this time period, if they tear apart, It, the ability of humans to actually be humans. I think we need to solve it…

MUSIC…

RAFFI:
Technology touches every aspect of our lives. This has always been true, and will be true forever. And if the explosion of technology over the past few decades has taught us anything, it’s that.. as tech changes… so does society. Social media, smartphones, even the internet itself… these things haven’t been around long at all, and today they dominate the way many of us work and live.

So, when it comes to AI, and this moment… when people talk about the oncoming revolution, or the massive changes that you and I and everyone should be prepared for… we know what this might mean… even if the future is totally unclear.

We called the show Technically Optimistic. But if you’ve listened, you know… we haven’t avoided talking about the downsides. We've discussed people’s worries and fears — even some of the most extreme ones.

But, we are *optimistic* because we believe in people. We believe that if people are curious about the technology all around us, and come to learn more about how it works, and ask big questions about tech for themselves… then we can broaden the conversations our society is having.

Because those conversations need more people in them — not just engineers and entrepreneurs. All kinds of people.

That way, we could all be better equipped to use technology for good… whether you're a teacher or a plumber.. a CEO, or a doctor.. an artist… or a parent.

But even more important… we need to be able to advocate for ourselves.

If we want someone to speak up when tech is biased against us, or when tech policy might be limiting human potential, rather than expanding it…..then that responsibility… is on us.

The person you're waiting for… is you.

Maria Ressa said it well… We've got to stop thinking of ourselves merely as users of technology… and start acting like what we are… Citizens.

And we believe people can do that. We believe you can do that. And that's why… we're technically optimistic.

MUSIC…

CREDITS

Technically Optimistic is produced by Emerson Collective, with original music by Mattie Safer.

We talked to tons of interesting people for this show, and they have plenty more to say about AI than what you heard in our episodes. We'll be back in the coming weeks with some bonus episodes—so look forward to going even deeper into the nuances of AI.

And, besides AI, there's a lot more I wanna talk about. We'll have new episodes coming soon– asking big questions about tech, humanity, and power… taking on issues like data privacy, democracy, and climate. So stay subscribed, and look out for season 2 later this year.

And we want to hear from you. What did you like about this series? What questions do you still have about AI? And are there big tech issues you want to hear us explore? Raach out to us at technically optimistic at emerson collective dot com.

If you're interested in learning more, or if you'd like transcripts of all our episodes, visit us on the web at emerson collective dot com, slash technically, dash optimistic, dash podcast. And follow us on social media, at emerson collective.

I'm Raffi Krikorian. Thanks again for listening. Technically Optimistic will be back soon.